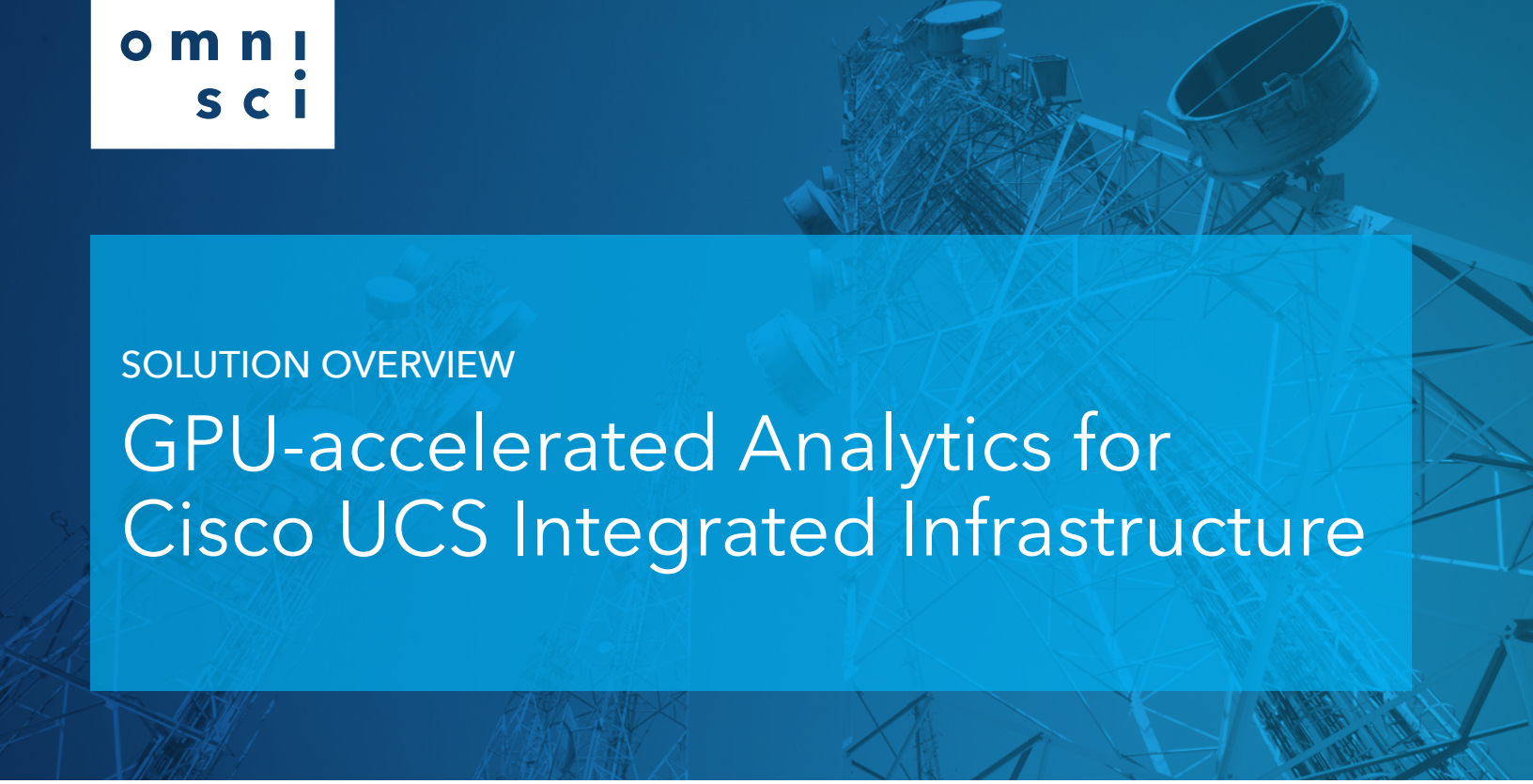


SOLUTION OVERVIEW

GPU-accelerated Analytics for Cisco UCS Integrated Infrastructure

Solution Highlights

High Performance with Linear Scalability

Cisco UCS Integrated Infrastructure for Big Data and Analytics is a simplified, intelligent infrastructure with the high performance and scalability that's needed to meet growing business demands. The OmniSci Core open source SQL engine was built for NVIDIA GPUs and can query many billions of rows in milli-seconds. Distributed, high-availability configurations take advantage of the Cisco Unified Computing System™ (Cisco UCS®) intelligent infrastructure.

Advantages of Cisco UCS Integrated Infrastructure

Cisco UCS Integrated Infrastructure for Big Data and Analytics is a proven platform for enterprise analytics applications with capabilities for powering analytics platforms accelerated by GPUs.

Ease of Deployment

Cisco UCS Manager simplifies infrastructure deploy-ment with an automated, policy-based mechanism that helps reduce configuration errors and system downtime. It offers easy scaling of the architecture with single- and multiple-rack deployments and proven, high performance linear scalability.

The OmniSci GPU-first architecture makes software deployment very straightforward: OmniSci requires no tuning before getting started.

Exceptional Performance at Scale

The OmniSci open source SQL engine takes the concept of distributed processing and extends it to the GPU, with not just one, but multiple GPUs per server and multiple servers per cluster. It redefines the limits on scale and speed with its ability to query tens of billions of records in milliseconds.

Interactive Visualization

Exceptional speed is useful for query results, but the OmniSci platform combined with the Cisco UCS Integrated Infrastructure for Big Data and Analytics also delivers similar speed to the end user with instantaneous visual analytics. While a lightning-fast SQL engine is valuable in its own right, being able to leverage that speed to interactively explore large datasets is truly transformative.

Instantly Query and Visually Analyze Data With the Power of GPUs

Big data systems are used in every industry. With this maturity come additional challenges in processing the vast amount of data quickly enough to be useful. The distributed nature of big data system processing means that one way of addressing the need for more processing power is to increase the number of CPUs and/or cores in each of the distributed servers.

Recent advances in Graphics Processing Units (GPUs) have extended their use beyond high-end computer gaming and supercomputing into the realms of data analytics and scientific computing. The raw processing power of a GPU server is orders of magnitude greater than that offered by CPU-based servers. GPUs are not general-purpose processors like CPUs. They work along with CPUs by offloading specific jobs (such as video analytics and deep learning) onto the GPUs, thereby accelerating those tasks and freeing CPUs for more general-purpose processing.

The OmniSci platform is designed for the most extreme analytic challenges. It harnesses the GPU's unique computing and visualization advantages so that users query and visualize billions of records in milliseconds. They create interactive dashboards displaying dozens of attributes that can correlate and cross-filter instantaneously. Whenever one filter is adjusted, all related charts, graphs, and maps redraw to match that context.

NVIDIA Tesla Accelerated Computing Platform

NVIDIA makes the world's fastest computing accelerators. Part of the NVIDIA Tesla Accelerated Computing Platform, Tesla GPU accelerators are built on the NVIDIA Kepler computing architecture and powered by Compute Device Unified Architecture (CUDA), a widely used parallel computing model.

This architecture makes these accelerators excellent for delivering record-setting acceleration and computing performance efficiency for a broad range of applications, including:

- Machine learning and data analytics;
- Seismic processing;
- Computational biology and chemistry;
- Weather and climate modeling;
- Image, video, and signal processing;
- Computational finance and physics;
- Computer-Aided Design (CAD); and
- Computational Fluid Dynamics (CFD).

The main features of the Tesla Accelerated Computing Platform include:

- Hyper-Q allows multiple CPU cores to simultaneously use the CUDA cores on single or multiple Kepler-based GPUs. This feature dramatically increases GPU utilization, simplifies programming, and decreases the amount of CPU idle time.

- Memory error protection meets a critical requirement for computing accuracy and reliability in data centers and supercomputing centers.

- Asynchronous transfer with dual Direct-Memory-Access (DMA) engines dramatically increases system performance by transferring data over the PCIe bus while the computing cores are processing other data.

- GPU Boost technology enables the end user to convert power headroom to higher clock speeds and achieve even greater acceleration for various high-performance computing workloads.

- The zero-power-idle feature increases data center energy efficiency by powering down idle GPUs when the system is running traditional, non-accelerated workloads.

Combining Cisco UCS® Integrated Infrastructure for Big Data and Analytics, GPU-accelerated hardware, and the OmniSci platform doesn't just change the way data is analyzed, it fundamentally changes the questions that can be asked and enables the kind of complex, iterative analysis that mirrors our own intuitive thought processes.

Cisco UCS Integrated Infrastructure for Big Data and Analytics

Organizations today must be sure that the underlying physical infrastructure can be deployed, scaled, and managed in a way that is agile enough to adapt to changing workloads and business requirements. Cisco UCS Integrated Infrastructure for Big Data and Analytics redefines the potential of the data center with its revolutionary approach to managing computing, network, and storage resources to successfully address the business needs of IT innovation and acceleration. Cisco UCS Integrated Infrastructure provides an end-to-end architecture for processing high volumes of structured and unstructured data for real-time processing, historical analysis, and archival purposes.

Cisco UCS 6300 Series Fabric Interconnects

Cisco UCS 6300 Series Fabric Interconnects provide high-bandwidth, low-latency connectivity for servers, with Cisco UCS Manager providing integrated, unified management for all connected devices. The Cisco UCS 6300 Series Fabric Interconnects are a core part of the Cisco UCS solution, providing low-latency, lossless 40 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and Fibre Channel functions. Cisco® fabric interconnects offer the full active–active redundancy, performance, and exceptional scalability needed to support the large number of nodes that are typical in clusters serving big data applications. Cisco UCS Manager enables rapid and consistent server configuration using service profiles and automates ongoing system maintenance activities, such as firmware updates across the entire cluster as a single operation. Cisco UCS Manager also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.

Cisco UCS C240 M5 Rack Server

The Cisco UCS C240 M5 rack servers are dualsocket, 2-rack-unit (2RU) servers offering industry leading performance and expandability for a wide range of storage and I/O-intensive big data and analytics workloads. These servers use the latest Intel® Xeon® Processor Scalable Family with up to 28 cores per socket. They support up to 24 Double-Data-Rate 4 (DDR4) dual in-line memory modules (DIMMs) for improved performance and lower power consumption. The DIMM slots are 3D XPoint ready, supporting next-generation nonvolatile memory technology. Depending on the server type, Cisco UCS rack servers offer a range of storage options. The Cisco UCS C240 M5 supports up to 24 Small Form-Factor (SFF) 2.5-inch drives (with support for up to 10 Non-Volatile Memory Express [NVMe].

Peripheral Component Interconnect Express [PCIe] solid-state drives [SSDs] on the (NVMe-optimized chassis version) or 12 large-form-factor (LFF) 3.5-inch drives plus 2 rear hot-swappable SFF drives with a Cisco 12-Gbps SAS Module RAID Controller. A modular LAN-on-Motherboard (mLOM) slot supports dual 40 Gigabit Ethernet network connectivity with the Cisco UCS Virtual Interface Card 1387.

Cisco UCS C480 M5 Rack Server

The Cisco UCS C480 M5 Rack Server is a quadsocket, 4-rack-unit (4RU) server offering a "no-compromise" balance of CPU, memory, storage, and I/O expansion in a 4RU form factor with the flexibility to scale. This server also uses the latest Intel Xeon Processor Scalable Family with up to 28 cores per socket. On the compute front, it has the ability to scale from two sockets to four using 2x (dual-socket) rack server modules. This server supports up to 24 front-accessible and up to 8 top-accessible SAS/SATA/SSD/NVM2/PCIe SSD-capable direct attached storage, bringing the total count of supported drives to 32. Support is provided for up to 24 DDR4 dual in-line memory modules (DIMMs) for improved performance.

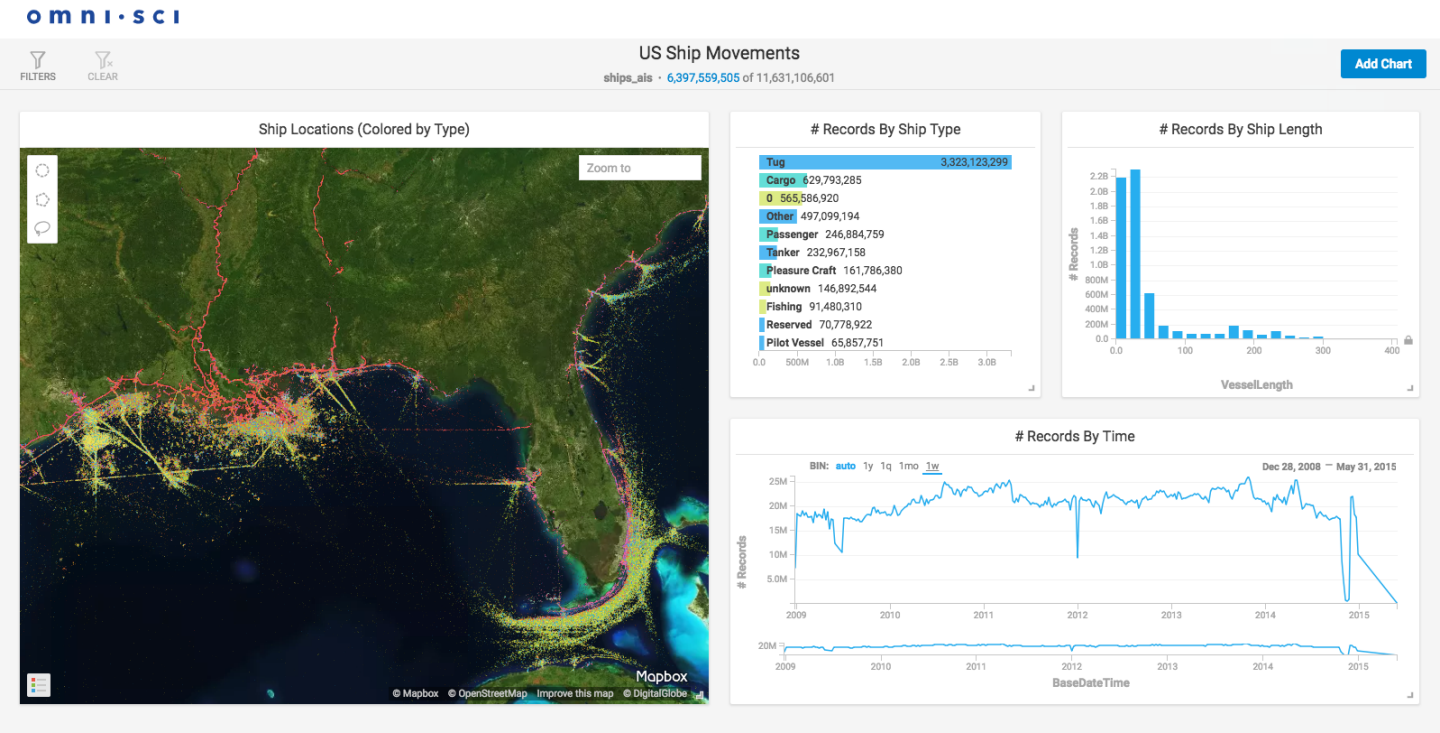
The OmniSci Platform

The OmniSci platform solves the problem of interactive performance at scale. Born out of research at MIT, OmniSci is a breakthrough analytics technology that harnesses the massive parallel processing and visual rendering power of graphics processing units (GPUs). Now analysts and data scientists can interactively analyze datasets with many billions of records and answer each question in milliseconds. OmniSci delivers an immersive, instantaneous,interactive way to explore massive datasets in real-time.

The open source OmniSci Core database allows analysts to use standard SQL and query extremely large datasets in milliseconds. With the OmniSci Low-Level Virtual Machine (LLVM) compilation engine, three-tier data caching, and parallel processing power, query results return in inter-active time, even as the data grows to billions of rows. OmniSci Core uses columnar storage, so every column is an index, eliminating the need for indexing data in order to optimize query performance.

OmniSci Core also includes a rendering engine custom built to use GPUs for in situ rendering of large amounts of data. This makes complex visualizations of geo charts and scatterplots both responsive and interactive, even with billions of data points or millions of polygons.

OmniSci Immerse is the visualization system that provides visual analytics designed to deliver the power of OmniSci processing and rendering to an interactive dashboard that analysts and data scientists use to explore the data. They can instantaneously cross-filter all the data features with interconnected maps, charts, and graphs that show their relationships. When the analyst changes location on a geo chart, all other visualizations refresh immediately in the context of that new location. Users can adjust the different data layers on a geo chart by changing their opacity, reordering them, or hiding them, as they interact with many metrics overlaid on the same map. Analysts can visually compare dozens of data sources in the same dashboard, and for fast-moving data streams, they can set visualization refresh intervals and watch very large datasets change in near-real-time.

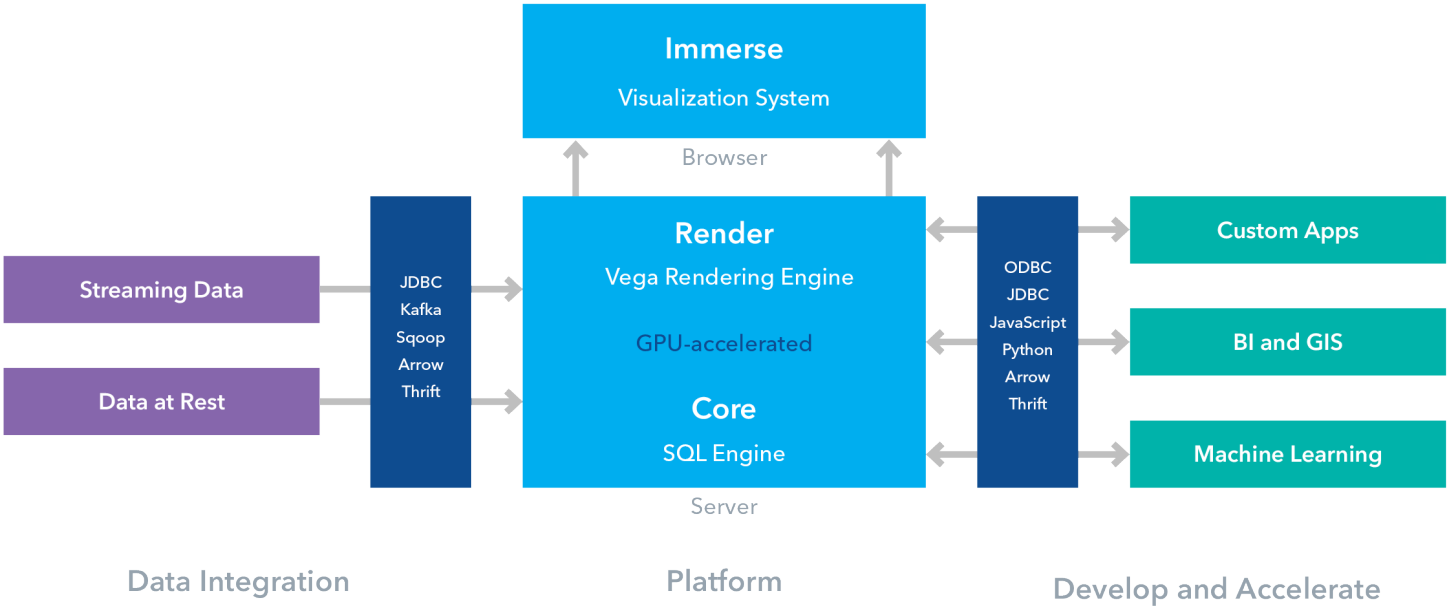


OmniSci Immerse combines GIS data and BI data into one dashboard, for a new breed of interactive geospatial exploration at scales beyond the reach of mainstream platforms.

The combined capabilities of OmniSci for the fastest SQL queries, big data rendering, and interactive visual analytics are enabling an entirely new form of analytics for a new type of continuous, operational decision-making. Until now, many use cases have been too large for mainstream analytic solutions to handle quickly enough.

Here are some of the industry challenges requiring fast SQL queries and interactive visualization on extremely large datasets:

- **Advertising**: Cross-filter millions of ad impressions.
- **Automotive**: Accelerate feature engineering for the artificial intelligence models that drive autonomous vehicles.
- **Capital Markets**: Analyze alternative datasets to create superior visibility into investment decisions.
- **Cybersecurity**: Speed cyber alert investigation and enable security operations teams to identify and repair vulnerabilities more quickly.
- **Defense**: Visualize the location of troops, vehicles, and weapons for military readiness and situational awareness.
- **Healthcare**: Model hospital staffing levels to deliver safe, efficient care.
- **Telecom**: Analyze millions of network logs per second to improve telecom network reliability.
- **Utilities**: See hundreds of millions of smart-meter records on a map to dynamically balance the electrical grid.

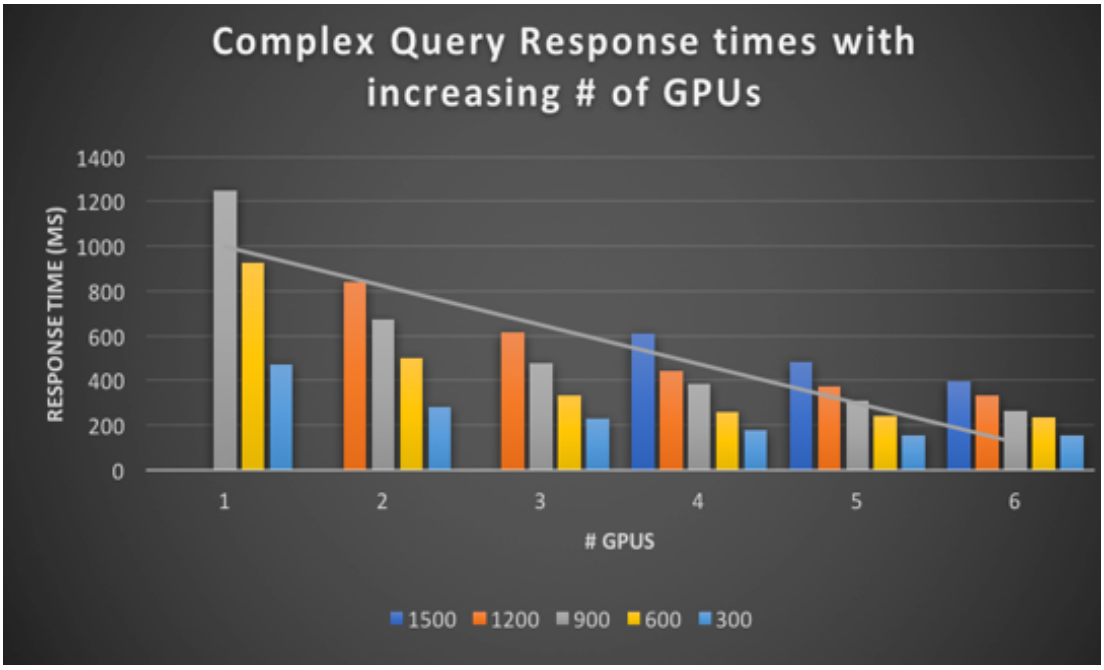


Scalable analytics solution with OmniSci platform powered by Cisco UCS Integrated Infrastructure

Performance Tests and Results

To showcase the performance of the OmniSci platform over Cisco UCS Integrated Infrastructure, multiple tests were performed with real telecommunications/mobile networks data. The testing methodology focused on data ingestion, fast query response times, and efficient utilization of the scale-up architecture provided by Cisco UCS C480 servers (supporting up to 4 CPUs and 6 GPUs).

These tests showcase the fast query response times compared to traditional database solutions. As shown in Figure 3, they also demonstrate the (near) linear performance relationship between the platform's speed and the number of GPUs.



Complex query response times with increasing number of GPUs

Reference Architecture

Basic Configuration	Advanced Configuration
8 Cisco UCS C240 M5, each with:	4 Cisco UCS C480 M5, each with:
• Dual Intel Xeon Processor Scalable Family 6132 CPUs (2 x 14 cores and 2.6 GHz) • 192 GB memory • 8 x 1.6-TB Enterprise Value SATA SSDs • 2 x P100 GPUs • Cisco UCS Virtual Interface Card 1387 with 2 x 40-Gbps port connectivity • Cisco 12-Gbps SAS Modular RAID Controller with 4-GB flash-based write cache • M.2 with 2 x 480-GB SSDs	• 4 x Intel Xeon Processor Scalable Family 6132 CPUs (2 x 14 cores and 2.6 GHz) • 768 GB memory • 16 x 1.6-TB Enterprise Value SATA SSDs • 6 x P100 GPUs • Cisco UCS Virtual Interface Card 1387 with 2 x 40-Gbps port connectivity • Cisco 12-Gbps SAS Modular RAID Controller with 4-GB flash-based write cache • M.2 with 2 x 480-GB SSDs

Conclusion

As data continues to accumulate from varied sources, business leaders must analyze that data in real-time to be able to draw actionable insights and make informed business decisions.

Cisco UCS Integrated Infrastructure for Big Data and Analytics, coupled with NVIDIA's GPU-accelerated hardware and the OmniSci platform solves the problem of interactive performance at scale. The solution harnesses the power of GPUs and provides a simplified, intelligent infrastructure with the scalability to meet growing business demands.

For More Information

For more on Cisco UCS big data solutions, see https://www.cisco.com/c/en/us/solutions/data-centervirtualization/big-data/index.html.

For more on Cisco UCS Integrated Infrastructure for Big Data, see https://blogs.cisco.com/datacenter/cpav5/.

For more on OmniSci, see https://www.omnisci.com.

For further information visit:
www.omnisci.com